ENTROPIC PRIORS FOR DISCRETE PROBABILISTIC NETWORKS AND FOR MIXTURES OF GAUSSIANS MODELS

C. C. RODRIGUEZ
Department of Mathematics and Statistics
University at Albany, SUNY
Albany NY 12222, USA
carlos@math.albany.edu
http://omega.albany.edu:8008/

Abstract. The ongoing unprecedented exponential explosion of available computing power, has radically transformed the methods of statistical inference. What used to be a small minority of statisticians advocating for the use of priors and a strict adherence to bayes theorem, it is now becoming the norm across disciplines. The evolutionary direction is now clear. The trend is towards more realistic, flexible and complex likelihoods characterized by an ever increasing number of parameters. This makes the old question of: *What should the prior be?* to acquire a new central importance in the modern bayesian theory of inference. Entropic priors provide one answer to the problem of prior selection. The general definition of an entropic prior has existed since 1988 [**?**], but it was not until 1998 [**?**] that it was found that they provide a new notion of complete ignorance. This paper re-introduces the family of entropic priors as minimizers of mutual information between the data and the parameters, as in [**?**], but with a small change and a correction. The general formalism is then applied to two large classes of models: Discrete probabilistic networks and univariate finite mixtures of gaussians. It is also shown how to perform inference by efficiently sampling the corresponding posterior distributions.

Key words: Bayesian Belief Networks, Mixture Models, Entropic Priors, Markov Chain Monte Carlo, MCMC, Generalized Inverse Gaussian distribution, Gamma Approximation to GIG

1. Introduction

Entropic Priors [**?**,**?**,**?**,**?**] minimize a type of mutual information between the data and the parameters [**?**]. Hence, Entropic Priors are the prior models that are most ignorant about the data. As Jaynes used to say: *they are maximally noncommittal with respect to missing information*. Entropic Priors (as opposed to other prior

assignments of probability) come with a guarantee: They include only the information in the likelihood, the initial guess, the hyper-parameter and the possible side conditions that are explicitly imposed, and nothing else. Entropic Priors provide a general recipe for prior probabilities that allow the enjoyment of the bayesian omelet even in high dimensional parameter spaces.

This paper presents the explicit computation of Entropic Priors for two classes of models: General Discrete Probabilistic Networks (a.k.a. Belief Nets, Bayesian Nets, BBNs) and for Mixtures of Gaussians Models. These models constitute the core of the probabilistic treatment of uncertainty in AI.

The paper is divided into 5 parts. Section 2, repeats the derivation in [**?**] (but with a small change and a correction) that Entropic Priors minimize mutual information between the data and the parameters. Section 3, presents the computation for discrete BBNs. Section 4 shows an application for classification. Section 5 computes the priors for the Mixture of Gaussians case. Finally some general remarks and conclusions are included in Section 6.

2. Entropic Priors are Most Ignorant Priors

Given a regular parametric hypothesis space, i.e. a Riemannian manifold of dominated probability distributions with volume element $g^{1/2}(\theta)d\theta$. Where $g(\theta)$ is the determinant of the Fisher information at θ. We denote by $f(x|\theta)$ the density (with respect to either Lebesgue or counting measure) of the distribution indexed by θ and by $\pi(\theta)$ a prior density on the parameters θ. The entropic prior is the π that makes the joint distribution

$$f(x_1, \ldots, x_\alpha, \theta) = \pi(\theta) \prod_{j=1}^{\alpha} f(x_j|\theta) \tag{1}$$

hardest to discriminate (in the sense of minimizing the Kullback number) from the independent model,

$$h(x_1, \ldots, x_\alpha) c g^{1/2}(\theta) \propto g^{1/2}(\theta) \left\{ \prod_{j=1}^{\alpha} h(x_j) \right\} \tag{2}$$

for a given fix density $h(x)$ on the data space. Where c is a normalization constant independent of θ and the x_js. Notice that $c > 0$ when the parameter space has finite volume. However, the solution to the optimization problem (5) (and hence, the entropic prior) does not depend on c and still makes sense for models with infinite volume. Notice further that the setting is coherent in the sense that the rhs of (2) is in fact proportional to the density of the model that assigns probabilities to the x_js according to h and independently of the θ which, according with (2), is uniform over the surface area of the model. This is true since Fisher information in the hypothesis space of α independent observations is α times the Fisher information in the hypothesis space of one observation and thus the volume element in the space of α observations is $\alpha^{k/2} g^{1/2}(\theta)$. i.e., the two volume elements are proportional and we assume the proportionality constant is included in c.

To simplify the notation let $x^\alpha = (x_1, \ldots, x_\alpha)$ and write,

$$I(\theta : h) = \int f(x|\theta) \log \frac{f(x|\theta)}{h(x)} dx \tag{3}$$

and

$$I(f\pi : hg^{1/2}) = \int f(x^\alpha|\theta)\pi(\theta) \log \frac{f(x^\alpha|\theta)\pi(\theta)}{h(x^\alpha)cg^{1/2}(\theta)} dx^\alpha d\theta \tag{4}$$

We have,

Theorem 1

$$\pi^* = \underset{\pi}{\operatorname{argmin}}\, I(f\pi : hg^{1/2}) \tag{5}$$

where the minimum is taken over all the proper priors on the parameter space, is given by the entropic prior:

$$\pi^*(\theta|\alpha, h) \propto e^{-\alpha I(\theta:h)} g^{1/2}(\theta) \tag{6}$$

Proof Using Fubbini's theorem, (1),(2) and the fact that π integrates to one, we can write

$$I(f\pi : hg^{1/2}) = \alpha \int \pi(\theta) I(\theta : h) d\theta + \int \pi(\theta) \log \frac{\pi(\theta)}{g^{1/2}(\theta)} d\theta - \log c. \tag{7}$$

Therefore using a Lagrange multiplier to enforce the normalization constraint ($\int \pi = 1$) we can find π^* by solving:

$$\underset{\pi}{\operatorname{argmin}} \int \left\{ \alpha\pi(\theta) I(\theta : h) + \pi(\theta) \log \frac{\pi(\theta)}{g^{1/2}(\theta)} + \lambda\pi(\theta) \right\} d\theta \tag{8}$$

Let $\mathcal{L}\,(\pi, \lambda)$ denote the expression inside the curly brackets in (8). The Euler-Lagrange equation for the optimal π^* is $\frac{\partial \mathcal{L}}{\partial \pi} = 0$ given by,

$$\alpha I + \log \pi^* - \log g^{1/2} + \lambda + 1 = 0. \tag{9}$$

From where we obtain the expression for the entropic prior given by (6).
Q.E.D.

2.1. BUT WHAT DOES IT MEAN?

First of all it needs to be clear that the above analysis is logically a priori. By this I mean that the actual numerical values of the observed data are not used, nor is the actual sample size number n of observed *i.i.d.* data vectors used. The parameters α and h of the entropic prior are the carriers of prior information. Notice also that, since the derivation was done on a *virtual* and not actual space of α observations, it makes sense to allow α to take non integer values as long as $\alpha > 0$. In fact an irrational α' is immediately obtained if we decide to change (in the final formula for the entropic prior) the entropy scale to *bits* by changing the original base of the logarithm in $I(\theta : h)$ from e to 2 so that $\alpha' = \alpha \log 2$. It is

however incorrect to claim that by starting the derivation with another base for the logarithm one will end up with a non integer α' as it was wrongly claimed in [**?**]. In fact the objective functions are proportional and they obviously produce the same π^*. To see the source of the mistake one just needs to notice that when the base of the log in (9) is 2 say, one has to exponentiate 2, and not e, in order to solve for π^*. This was first pointed out to me by Ariel Caticha, who then tried to build a justification for an entropic prior with fix $\alpha = 1$ in [**?**].

2.1.1. *Imaginary* α

Allowing α to be not just a real number but a Clifford number, in particular to be a pseudo scalar, opens up a garden of unexplored possibilities. This may not be as insane as it first appears to be, if one thinks of the resulting prior as the density of a Clifford valued probability measure (see [**?**]). Moreover, if I (entropy) could be justified as S (action) then the resulting *prior* $e^{iS/\hbar}$ (relative to local ignorance) would take a familiar form. Going with the flow of this (for now) applied numerology this would point to current physical theory to be based on the order of 10^{66} equivalent a priori observations! (i.e. expressing $\hbar$ in geometrized units).

2.2. RECIPES FOR CHOOSING α AND h

The values of the hyperparameters α and h of the entropic prior need to be fixed in order to obtain numerical assignments of probabilities. To fix h we need to specify a function (i.e. an a priori density $h(x)$ for the data) which involves, in principle, the specification of an infinite number of parameters. Nevertheless, the importance of the a priori biases introduced by h are modulated by the value of the real positive parameter α. Take α sufficiently close to 0 and the prior will be blind to the specific form of h and controlled by the volume element $g^{1/2}d\theta$ (i.e. uniform over the model surface, see [**?**]). There is a close similarity with the problem of choosing a kernel and a bandwidth in density estimation. As it is the case in density estimation, the specific form of the kernel is not as critical as the choice of the smoothness parameter. A natural choice for h is to use $h(x) = f(x|\theta_0)$ where θ_0 is the best current guess for the value of θ. If we assume the value of θ_0 to be unknown then we can consider the entropic prior model, which is now indexed by the $1 + k$ parameters (α, θ_0), to be another regular hypothesis space that needs a prior on its parameters. The entropic prior on the entropic prior, on the entropic prior,. . ., etc is, in principle, computable. The possibility of a chain of entropic priors for α was first given to first level in [**?**] and for all levels in [**?**]. Another general alternative is to use the empirical bayes approach (see [**?**]). Finally, just fixing α to an arbitrary small value (≈ 1) and using $\hat{\theta_0}$ the mle (maximum likelihood estimator) or MAP (Maximum A posteriori Probability), with an easy to handle conjugate prior, for θ has been shown to perform well in simulation experiments.

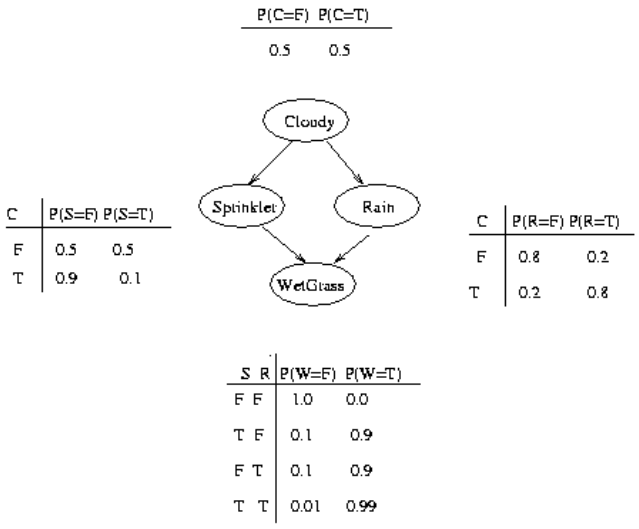

Figure 1. DAG for the Sprinkler Problem

3. The Entropic Prior of a Discrete Probabilistic Network

An understanding of Cox's [**?**] argument should be sufficient to impose the rules of probability to the treatment of uncertainty in AI. But it has taken, however, a long heated debate (see [**?**] and [**?**]), the invention of new efficient methods of computation (e.g. the junction tree algorithm, see [**?**]) and the publication of Pearl's text [**?**], to arrive at today's dominant view of a complete probabilistic approach.

3.1. DAGS

The current recipe for the thinking machine consists of a fully bayesian probabilistic treatment of a long vector of facts (the data). The main approach for encoding prior information about an specific domain of application, is not the prior, but the likelihood. An a priori network of conditional independence assumptions is typically provided by means of a Directed Acyclic Graph (DAG) that is supposed to encode an expert's knowledge of causal relations among observable facts.

The canonical textbook example is displayed in fig 1. The arrows indicate causality. Thus, the presence of the arrow from *Cloudy* to *Rain* represents the fact that the sky being cloudy is a possible cause for rain. More important is the absence of arrows which indicate independence. Thus, the picture shows that conditionally on the values of *Sprinkler* and *Rain*, *Cloudy* is independent of *WetGrass*. The entries of the tables of conditional probabilities constitute the parameters of the

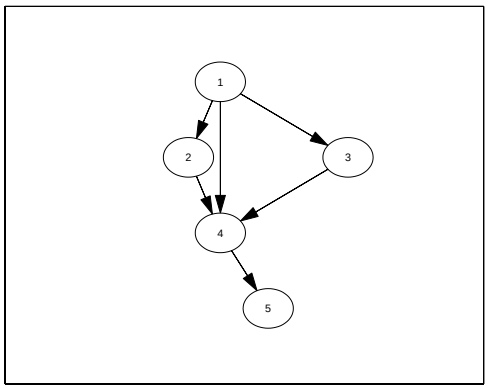

Figure 2. Example of a DAG

DAG. In the case of fig 1 there are 9 independent parameters. We can think of a DAG as a convenient way to specify a high dimensional submanifold of the space of all joint distributions of the variables under consideration. For example, the pictured DAG (with unspecified tables) represents a 9 dimensional submanifold of the 15 dimensional simplex of all the assignments of probability on the $2^4 = 16$ possible observations of the binary variables (C, S, R, W). The DAG in fig 1 specifies the joint distribution of all the variables (C, S, R, W) in terms of the parameters θ (i.e. table entries) as,

$$P(C, R, S, W) = P(C)P(R|C)P(S|R)P(W|R, S). \tag{10}$$

Each of the factors on the right of equation (10) can be read off the tables provided in fig 1. For example,

$$P(C = T, R = T, S = F, W = F) = (0.5)(0.8)(0.5)(0.1) = 0.02 \tag{11}$$

In order to provide general formulas for DAGs we number the vector of variables by $x = (x_1, x_2, x_3, x_4) = (C, R, S, W)$ and parameterized the joint distribution with a vector θ of parameters as in,

$$\theta_{4w}(r, s) = P(W = w|R = r, S = s) \tag{12}$$

Thus, labeling $F = 1$ and $T = 2$, (11) becomes,

$$P(2, 2, 1, 1|\theta) = \theta_{12}\theta_{22}(2)\theta_{31}(2)\theta_{41}(2, 1) \tag{13}$$

3.2. WHO IS WHO ON A DAG: GENERAL NOTATION

This section provides some definitions and notations that are needed for writing the entropic prior on a general DAG. All the examples refer to fig 2.

Directed Graph: An ordered pair (V, E) where V is a set of vertices (e.g. $V = \{1, 2, 3, 4, 5\}$) and $E \subset V \times V$ is a set of directed edges. e.g.,

$$E = \{(1,2), (1,3), (1,4), (2,4), (3,4), (4,5)\}$$

DAG: A Directed Acyclic Graph is a directed graph without cycles. (e.g. fig 2).
Parents: $\text{pa}(k)$ denotes the set of parents for the vertices $k \in V$. (e.g. $\text{pa}(1) = \phi, \text{pa}(5) = \{4\}, \text{pa}(4) = \{1, 2, 3\}$).
Ancestors: $\text{an}(k)$ denotes the set of ancestors of $k \in V$. (e.g. $\text{an}(2) = \{1\}, \text{an}(5) = \{1, 2, 3, 4\}, \text{an}(1) = \phi$). Clearly,

$$\text{an}(k) = \text{pa}(k) \bigcup_{j \in \text{pa}(k)} \text{an}(j) \tag{14}$$

Ancestors that are not Parents: Denoted by $\text{ap}(k)$

$$\text{ap}(k) = \text{an}(k) \setminus \text{pa}(k) \tag{15}$$

(e.g. $\text{ap}(5) = \{1, 2, 3\}, \text{ap}(4) = \phi, \text{ap}(2) = \phi$).
Notation:

$$x_{\text{pa}(k)} \equiv \{x_j : j \in \text{pa}(k)\} \tag{16}$$

e.g.

$$x_{\text{pa}(1)} = \phi, \; x_{\text{pa}(4)} = \{x_1, x_2, x_3\}$$

Notation: $\sum_{x_{\text{pa}(k)}}$ denotes the multiple sum over all the possible values of the variables that are parents of vertice $k \in V$. e.g.

$$\sum_{x_{\text{pa}(4)}} \equiv \sum_{x_1} \sum_{x_2} \sum_{x_3}$$

The notation introduced with equation (13) generalizes naturally for any number of discrete variables. Given a DAG with set of vertices V we let $x = \{x_k : k \in V\}$. Hence, the joint distribution of the variables of a given DAG is given by,

$$\begin{aligned} p(x|\theta) &= \prod_{k \in V} p(x_k | x_{\text{pa}(k)}, \theta) \\ &= \prod_{k \in V} \theta_{k x_k}(x_{\text{pa}(k)}) \end{aligned} \tag{17}$$

We are now ready to compute.

3.3. ENTROPY OF A DAG

Given a DAG, the Kullback number between two sets of parameters θ and μ is,

$$I(\theta : \mu) = E_\theta \left[\log \frac{p(x|\theta)}{p(x|\mu)} \right] \tag{18}$$

Using (17) and interchanging expectation with summation we obtain,

$$I(\theta:\mu) = \sum_{k\in V} E_\theta\left[\log\frac{\theta_{kx_k}(x_{\mathrm{pa}(k)})}{\mu_{kx_k}(x_{\mathrm{pa}(k)})}\right] \tag{19}$$

Now for each $k \in V$ compute the unconditional expectation in (19) by first conditioning on the values of $x_{\mathrm{pa}(k)}$ to obtain,

$$\begin{aligned} E_\theta\left[\log\frac{\theta_{kx_k}(x_{\mathrm{pa}(k)})}{\mu_{kx_k}(x_{\mathrm{pa}(k)})}\middle| x_{\mathrm{pa}(k)}\right] &= \sum_{j=1}^{r_k}\theta_{kj}(x_{\mathrm{pa}(k)})\log\frac{\theta_{kj}(x_{\mathrm{pa}(k)})}{\mu_{kj}(x_{\mathrm{pa}(k)})} \\ &= I(\theta_k(x_{\mathrm{pa}(k)}):\mu_k(x_{\mathrm{pa}(k)})) \end{aligned} \tag{20}$$

where the last equality is a definition and it was assumed that x_k can take r_k discrete values. Taking expectations over the $x_{\mathrm{pa}(k)}$ and replacing in (19) we obtain,

$$I(\theta:\mu) = \sum_{k\in V}\sum_{x_{\mathrm{pa}(k)}} p(x_{\mathrm{pa}(k)}|\theta)\; I\left(\theta_k(x_{\mathrm{pa}(k)}):\mu_k(x_{\mathrm{pa}(k)})\right). \tag{21}$$

Finally, using the fact that,

$$\begin{aligned} p(x_{\mathrm{pa}(k)}|\theta) &= \sum_{x_{\mathrm{ap}(k)}} p(x_{\mathrm{ap}(k)}, x_{\mathrm{pa}(k)}|\theta) \\ &= \sum_{x_{\mathrm{ap}(k)}}\prod_{j\in\mathrm{an}(k)} p(x_j|x_{\mathrm{pa}(j)},\theta) \\ &= \sum_{x_{\mathrm{ap}(k)}}\prod_{j\in\mathrm{an}(k)} \theta_{jx_j}(x_{\mathrm{pa}(j)}) \end{aligned} \tag{22}$$

we obtain the expression for the entropy,

$$I(\theta:\mu) = \sum_{k\in V}\sum_{x_{\mathrm{pa}(k)}}\left\{\sum_{x_{\mathrm{ap}(k)}}\prod_{j\in\mathrm{an}(k)}\theta_{jx_j}(x_{\mathrm{pa}(j)})\right\} I(\theta_k(x_{\mathrm{pa}(k)}):\mu_k(x_{\mathrm{pa}(k)})). \tag{23}$$

Thus, formula (21) shows that the total entropy for a DAG is obtained by adding the entropies for each node. The entropy of a node is computed as an average of all the possible entropies obtained for the different values of the parents of that node. In practice formula (23) may be too expensive to compute and it may be necessary to use a Monte Carlo estimate.

3.4. VOLUME ELEMENT OF A DAG

To compute the Fisher metric, write θ as a long vector and use the fact (see [**?**]) that,

$$I(\theta:\theta+\epsilon v) = \frac{\epsilon^2}{2}\sum_{i,j} g_{ij}(\theta)v^i v^j + o(\epsilon^2) \tag{24}$$

It then follows immediately from (23) that the Fisher matrix is block diagonal. Each block corresponds to the $(r_k - 1) \times (r_k - 1)$ (Fisher matrix $G_k(\theta_k(x_{\mathrm{pa}(k)}))$ associated to the kth node, multiplied by the scalar $p(x_{\mathrm{pa}(k)}|\theta)$. The determinant, $g(\theta)$, of the Fisher matrix is then given by the product of the determinants of each of the blocks. We have

$$g(\theta) = \prod_{k \in V} \prod_{x_{\mathrm{pa}(k)}} \left\{ \sum_{x_{\mathrm{ap}(k)}} \prod_{j \in \mathrm{an}(k)} \theta_{j x_j}(x_{\mathrm{pa}(j)}) \right\}^{r_k - 1} \det G_k \left(\theta_k(x_{\mathrm{pa}(k)}) \right) \tag{25}$$

Finally using the fact that G_k is the Fisher matrix of a multinomial with parameters $\theta_{k1}(x_{\mathrm{pa}(k)}), \ldots, \theta_{k r_k}(x_{\mathrm{pa}(k)})$ we have,

$$\det G_k \left(\theta_k(x_{\mathrm{pa}(k)}) \right) = \frac{1}{\prod_{j=1}^{r_k} \theta_{kj}(x_{\mathrm{pa}(k)})} \tag{26}$$

replacing (26) in (25) and taking square root we obtain the expression for the volume element,

$$g^{1/2}(\theta)\, d\theta = \prod_{k \in V} \prod_{x_{\mathrm{pa}(k)}} \frac{\left\{ \sum_{x_{\mathrm{ap}(k)}} \prod_{j \in \mathrm{an}(k)} \theta_{j x_j}(x_{\mathrm{pa}(j)}) \right\}^{(r_k - 1)/2}}{\prod_{j=1}^{r_k} \theta_{kj}^{1/2}(x_{\mathrm{pa}(k)})}\, d\theta \tag{27}$$

3.5. THE ENTROPIC PRIOR FOR A DAG

To obtain (6) we use (23), (22) and (27) to get,

$$\begin{aligned} \pi(\theta|\alpha, \mu) \propto \quad & \prod_{k \in V} \prod_{x_{\mathrm{pa}(k)}} \left\{ \frac{p^{(r_k - 1)}(x_{\mathrm{pa}(k)}|\theta)}{\prod_{j=1}^{r_k} \theta_{kj}(x_{\mathrm{pa}(k)})} \right\}^{1/2} \\ & \exp \left\{ -\alpha p(x_{\mathrm{pa}(k)}|\theta) I(\theta_k(x_{\mathrm{pa}(k)}) : \mu_k(x_{\mathrm{pa}(k)})) \right\} \end{aligned} \tag{28}$$

3.6. POSTERIOR

Let us assume that there is available a set of N independent observations

$$D = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\} \tag{29}$$

where each $x^{(t)} = (x_1^{(t)}, \ldots, x_n^{(t)})$ is an $|V| = n$ dimensional vector containing the observed values of the nodes of a general DAG (V, E). As usual the posterior is given by Bayes theorem as,

$$\pi(\theta|D, \alpha, \mu) \propto f(D|\theta)\ \pi(\theta|\alpha, \mu) \tag{30}$$

where the likelihood is given by,

$$\begin{aligned} f(D|\theta) &= \prod_{t=1}^{N} f(x^{(t)}|\theta) \\ &= \prod_{t=1}^{N}\prod_{k=1}^{n} f(x_k^{(t)}|x_{\mathrm{pa}(k)}^{(t)}, \theta) \\ &= \prod_{t=1}^{N}\prod_{k=1}^{n} \theta_{kx_k^{(t)}}(x_{\mathrm{pa}(k)}^{(t)}) \end{aligned} \tag{31}$$

Let us partition the set of vertices into two groups, those with parents and those without (orphans). For the orphan nodes, i.e. for $k \in V$ such that $\mathrm{pa}(k) = \phi$ and for $i = 1, 2, \ldots, r_k$ define

$$n_{ki}(\phi) = \left|\{t : x_1^{(t)} = i\}\right| \tag{32}$$

and for $k \in V$ with $\mathrm{pa}(k) \neq \phi$ and $i = 1, 2, \ldots, r_k$

$$n_{ki}(x_{\mathrm{pa}(k)}) = \left|\{t : x_k^{(t)} = i \text{ and } x_{\mathrm{pa}(k)}^{(t)} = x_{\mathrm{pa}(k)}\}\right| \tag{33}$$

Replacing these counts into (31) we obtain,

$$f(D|\theta) = \prod_{k=1}^{n}\prod_{x_{\mathrm{pa}(k)}}\prod_{i=1}^{r_k}\{\theta_{ki}(x_{\mathrm{pa}(k)})\}^{n_{ki}(x_{\mathrm{pa}(k)})} \tag{34}$$

To simplify the notation let us write simply by p_k the expression (22) which is always a probability that depends only on the ancestors of the node k. Let us also just write $\theta_{ki}, n_{ki}, \mu_{ki}$ instead of $\theta_{ki}(x_{\mathrm{pa}(k)}), \ldots$ and keep implicit their dependence on given values of the parents. With this notation the posterior becomes,

$$\pi(\theta|D, \alpha, \mu) \propto \prod_{k\in V}\prod_{x_{\mathrm{pa}(k)}} p_k^{(r_k-1)/2}\prod_{i=1}^{r_k}\left\{\theta_{ki}^{n_{ki}-\frac{1}{2}}\ \exp\left(-\alpha p_k\theta_{ki}\log\frac{\theta_{ki}}{\mu_{ki}}\right)\right\} \tag{35}$$

were we have used (20) to write the exponential in (28) as a product of r_k factors.

4. Example: Naïve Bayes

When the DAG has the form shown in fig 3 the general formulas have simpler forms. This case is known as naïve bayes and it is often used as an approximation

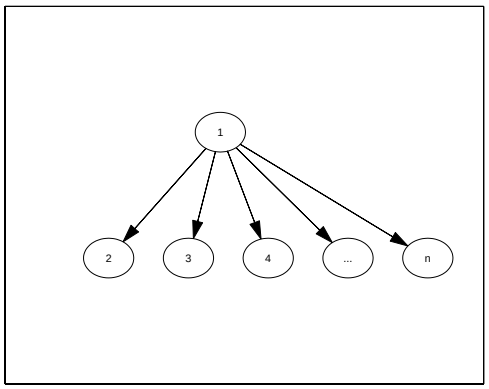

Figure 3. DAG for Naïve Bayes

in discrimination problems. For this case, $V = \{1, \ldots, n\}$, $\mathrm{pa}(1) = \phi$, and for $k \neq 1$ we have $\mathrm{pa}(k) = \{1\}, \mathrm{an}(k) = \{1\}, \mathrm{ap}(k) = \phi$ and,

$$p(x_{\mathrm{pa}(k)}|\theta) = p(x_1|\theta_1) = \theta_{1x_1} \tag{36}$$

The expression for the entropy (23) becomes,

$$I(\theta : \mu) = I(\theta_1, \mu_1) + \sum_{k=2}^{n} \sum_{j=1}^{r_1} \theta_{1j}\ I(\theta_k(j) : \mu_k(j)) \tag{37}$$

and the volume element (27) reduces to,

$$g^{1/2}(\theta)\ d\theta = \frac{\left(\prod_{j=1}^{r_1} \theta_{1j} \right)^{(\sum_{k=2}^{n} r_k - 1)/2}}{\left(\prod_{j=1}^{r_1} \prod_{k=2}^{n} \prod_{i=1}^{r_k} \theta_{ki}(j) \right)^{1/2}}\ d\theta \tag{38}$$

The entropic prior is then easily computed by multiplying $\exp(-\alpha I(\theta : \mu))$ (obtained from (37)) by (38).

4.1. POSTERIOR

For naïve bayes the likelihood is given by,

$$f(D|\theta) = \prod_{i=1}^{N} \theta_{1x_1^{(i)}} \prod_{k=2}^{n} \theta_{kx_k^{(i)}}(x_1^{(i)}) \tag{39}$$

Replacing the counts into (39) we obtain,

$$f(D|\theta) = \left(\prod_{j=1}^{r_1} \theta_{1j}^{n_{1j}}\right) \left(\prod_{j=1}^{r_1}\prod_{k=2}^{n}\prod_{i=1}^{r_k} (\theta_{ki}(j))^{n_{ki}(j)}\right) \tag{40}$$

Letting,

$$m = \frac{1}{2}\left(\sum_{k=2}^{n} r_k - n\right) \tag{41}$$

we can write the posterior as,

$$\pi(\theta|D,\alpha,\mu) \propto \left\{\prod_{j=1}^{r_1} \theta_{1j}^{m+n_{1j}-\alpha\theta_{1j}} \exp\left(-(\alpha \log\frac{1}{\mu_{1j}})\theta_{1j}\right)\right\} \tag{42}$$

$$\left\{\prod_{j=1}^{r_1}\prod_{k=2}^{n}\prod_{i=1}^{r_k} (\theta_{ki}(j))^{n_{ki}(j)-\frac{1}{2}-\alpha\theta_{1j}\theta_{ki}(j)} \exp\left(-(\alpha\theta_{1j}\log\frac{1}{\mu_{ki}(j)})\theta_{ki}(j)\right)\right\}$$

4.2. THE ENTROPIC SAMPLER

A combination of Gibbs and Metropolis can be used for sampling the posterior (42). The parameters are naturally grouped in blocks θ_k, where,

$$\begin{aligned} \theta_k &= \theta_k(x_{\mathrm{pa}(k)}) \\ &= (\theta_{k1},\ldots,\theta_{kr_k}) \text{ with } \sum_{i=1}^{r_k}\theta_{ki} = 1 \end{aligned} \tag{43}$$

are distributed over the simplex of dimension $r_k - 1$. It can be readily seen from (42) that the marginal joint distributions of the θ_k blocks are all of the generic form,

$$f(y_1, y_2, \ldots, y_{r-1}) \propto \prod_{j=1}^{r} \left\{y_j^{\alpha_j - 1}\ e^{-\beta_j y_j}\right\} \tag{44}$$

with $y_j \geq 0$ and $y_r = 1 - \sum_{j=1}^{r-1} y_j$. The parameters α_j and β_j are different for the parent node and for the children nodes. For the parent,

$$\alpha_j = 1 + m + n_{1j} - \alpha\theta_{1i} \approx 1 + m + n_{1j} \tag{45}$$

$$\beta_j = \alpha\left(\log\frac{1}{\mu_{1j}} + \sum_{k=2}^{n} I(\theta_k(j) : \mu_k(j))\right) \tag{46}$$

For the children blocks the parameters are,

$$\alpha_j = n_{ki}(j) + \frac{1}{2} - \alpha\theta_{1j}\theta_{ki}(j) \approx n_{ki}(j) + \frac{1}{2} \tag{47}$$

$$\beta_j = \alpha\theta_{1j}\log\frac{1}{\mu_{ki}(j)} \tag{48}$$

Excellent initial distributions for Metropolis are obtained by using the following,

Lemma 1 *Let $y_1, y_2, \ldots, y_r$ be independent with y_j following a Gamma distribution with parameters (α_j, β_j). Let,*

$$z_j = \frac{y_j}{y_1 + \cdots + y_r} \text{ for } j = 1, \ldots, r-1 \tag{49}$$

then the joint density of the z_j's is given by,

$$f(z_1, \ldots, z_{r-1}) \propto \frac{\prod_{j=1}^{r} z_j^{\alpha_j - 1}}{\left(\sum_{j=1}^{r} \beta_j z_j\right)^{\alpha_1 + \alpha_2 + \cdots + \alpha_r}} \tag{50}$$

where $z_r \equiv 1 - z_1 - z_2 - \cdots - z_{r-1}$.

Proof Notice that (50) is a generalization of the classic result for the Dirichlet distribution obtained when all the β_j's are equal, in which case the denominator becomes proportional to 1. To prove (50) just condition on $y_r = y$ so that the transformation (49) from the y_j's to the z_j's for $j = 1, 2, \ldots, r-1$ is one to one with inverse,

$$y_j = \frac{y z_j}{z_r} \text{ for } j = 1, \ldots, r-1 \tag{51}$$

To show (51) just notice that,

$$\begin{aligned} y\ z_j &= \frac{y_j\ y}{y + \sum_{i=1}^{r-1} y_i} && (52)\\ &= y_j \left(1 - \frac{\sum_{i=1}^{r-1} y_i}{y + \sum_{i=1}^{r-1} y_i}\right) && (53)\\ &= y_j \left(1 - \sum_{i=1}^{r-1} z_i\right) && (54)\\ &= y_j\ z_r && (55) \end{aligned}$$

where we have used (49) and the definition of z_r. The probability density of observing $z_1, \ldots, z_{r-1}$ is then,

$$f(z_1, \ldots, z_{r-1}) = \int_0^\infty f(z_1, \ldots, z_{r-1} | y_r = y) g_r(y)\ dy \tag{56}$$

where g_j for $j = 1, \ldots, r$ are the gamma densities of the y_j. Using the definition of the z_j's given in (49), the assumed independence of the y_j's, and the change of variables theorem together with (62), we have,

$$f(z_1, \ldots, z_{r-1} | y_r = y) = \left(\prod_{j=1}^{r-1} g_j\left(\frac{y\ z_j}{z_r}\right)\right)\ \frac{1}{z_r}\left(\frac{y}{z_r}\right)^{r-1} \tag{57}$$

The expression outside the product is the determinant of the Jacobian of the transformation (51). This can be seen by noticing that the Jacobian matrix is,

$$J = \frac{y}{z_r^2}\begin{bmatrix} z_1+z_r & z_1 & \dots & z_1 \\ z_2 & z_2+z_r & \dots & z_2 \\ & & \ddots & \\ z_{r-1} & z_{r-1} & \dots & z_{r-1}+z_r \end{bmatrix} \tag{58}$$

and compute its determinant by subtracting from each column the column that follows, to obtain,

$$\det J = \left(\frac{y}{z_r^2}\right)^{r-1}\begin{vmatrix} z_r & 0 & \dots & 0 & z_1 \\ -z_r & z_r & \dots & 0 & z_2 \\ & & \ddots & & \\ 0 & 0 & \dots & -z_r & z_{r-1}+z_r \end{vmatrix} \tag{59}$$

and expanding along the last column,

$$\begin{aligned} \det J &= \left(\frac{y}{z_r^2}\right)^{r-1}\left(z_1 z_r^{r-2} + z_2 z_r^{r-2} + \ldots + z_{r-2}z_r^{r-2} + (z_{r-1}+z_r)z_r^{r-2}\right) \\ &= \left(\frac{y}{z_r^2}\right)^{r-1} z_r^{r-2} \\ &= \frac{1}{z_r}\left(\frac{y}{z_r}\right)^{r-1} \end{aligned} \tag{60}$$

This proves (57). Replacing (57) into (56) and using the expressions for the gamma densities we obtain,

$$f(z_1,\ldots,z_{r-1}) \propto \left(\prod_{j=1}^{r} z_j^{\alpha_j-1}\right)\frac{1}{z_r}\int_0^\infty \left(\frac{y}{z_r}\right)^{\sum_{j=1}^r \alpha_j - 1} \exp\left\{-\left(\frac{1}{z_r}\sum_{j=1}^r \beta_j z_j\right) y\right\} dy \tag{61}$$

this is a simple gamma integral. Integrating out and simplifying the z_r's we obtain the desired result (50).

Q.E.D.

To generate approximate samples from (44) we use the Lemma but with $\tilde{\beta}_j$ chosen so that,

$$\frac{C}{\left(\sum_{j=1}^r \tilde{\beta}_j z_j\right)^{\alpha_1+\cdots+\alpha_r}} \approx \exp\left(-\sum_{j=1}^r \beta_j z_j\right) \tag{62}$$

where the constant C does not depend on the z_j. To find the $\tilde{\beta}_j$ just write the left side of (62) in exponential form and use,

$$\log(\tilde{\beta}_1 z_1 + \cdots + \tilde{\beta}_r z_r) = \log(\tilde{\beta}_r) + \log\left(1 + \frac{\tilde{\beta}_1 - \tilde{\beta}_r}{\tilde{\beta}_r} z_1 + \cdots + \frac{\tilde{\beta}_{r-1} - \tilde{\beta}_r}{\tilde{\beta}_r} z_{r-1}\right) \tag{63}$$

together with,

$$\log(1 + z) = z + o(z) \tag{64}$$

we obtain, that in order for (62) to be true, we must have,

$$\frac{\tilde{\beta}_j - \tilde{\beta}_r}{\tilde{\beta}_r} \sum_{i=1}^{r} \alpha_i = \beta_j - \beta_r \tag{65}$$

we can then use,

$$\tilde{\beta}_r = \sum_{i=1}^{r} \alpha_i \tag{66}$$

$$\tilde{\beta}_i = \beta_i - \beta_r + \tilde{\beta}_r \tag{67}$$

Metropolis corrections are needed to correct for the approximations introduced in (45), (47) and (64).

4.3. TEST: CREDIT CARD CLASSIFICATION EXAMPLE

We tested the performance of the MCMC sampler on a standard set of 10000 data records containing the 13 variables in table 1. Most of the node names are self

Nodes	Sizes
Card = C	2(4)
Gender = G	2
Country = Y	3
Age = A	9
State = S	13
Education = E	5
Marital = M	2
Occupation = O	5
Total children = T	6
Income = I	8
House owner = H	2
Cars owned = R	5
Children home = N	6

TABLE 1. Data Records in Example

explanatory. Card, originally contained the type of credit card owned by the individual with categories: *no card, regular, gold, platinum*. These were later reduced to only two categories: {*no card, regular*} *and* {*gold, platinum*}. The data contains individuals from the three north american countries: *Mexico, US, Canada*. However, the majority of records are from the US. The *Children home* variable contains information about the actual number of children living at home with the individual.

4.3.1. *The Bayes Classifier*
To test the performance of the entropic sampler we chose at random 100 individuals to be used as the observed data and 1000 to test the bayes classifier. The bayes classifier simply assigns the category with highest posterior probability.

Let D be the observed $N = 100$ records and let $x_2, \ldots, x_n$ (here $n = 13$) be the values of all the nodes except the first (i.e. *Card*) for an individual that we want to classify. The bayes classifier allocates $x_1 = 1$ if,

$$P\left(x_1 = 1 | x_2, \ldots, x_n, D\right) > P\left(x_1 = 2 | x_2, \ldots, x_n, D\right) \tag{68}$$

we compute both sides with,

$$\begin{aligned} P\left(x_1 = j | x_2, \ldots, x_n, D\right) &= \int P\left(x_1 = j, \theta | x_2, \ldots, x_n, D\right)\ d\theta \\ &\propto \int P\left(x_1 = j, x_2, \ldots, x_n, \theta | D\right)\ d\theta \\ &= \int P\left(x_1 = j, x_2, \ldots, x_n | \theta\right)\ \pi(\theta | D)\ d\theta \end{aligned} \tag{69}$$

where we have assumed that the values of the individual to be classified are independent of the observed data D. We use the MCMC sampler to estimate (69) for $j = 1$ and $j = 2$. Thus, if the sampler produces $\theta^{(1)}, \ldots, \theta^{(M)}$ samples from the posterior $\pi(\theta|D)$ we classify $x_1 = 1$ if,

$$\sum_{t=1}^{M} p(1, x_2, \ldots, x_n | \theta^{(t)}) > \sum_{t=1}^{M} p(2, x_2, \ldots, x_n | \theta^{(t)}) \tag{70}$$

To avoid underflows it is better to use only ratios. A more stable rule is then: assign $x_1 = 1$ if,

$$\sum_{t=1}^{M} \left(1 - \frac{p(2, x_2, \ldots, x_n | \theta^{(t)})}{p(1, x_2, \ldots, x_n | \theta^{(t)})}\right) \frac{p(1, x_2, \ldots, x_n | \theta^{(t)})}{p(1, x_2, \ldots, x_n | \theta^{(1)})} > 0 \tag{71}$$

4.3.2. *Preliminary Results*
Table 2 shows the results of running the sampler with different parameter values. The burn column contains the number of complete sweeps performed and discarded before collecting samples. The other columns are: M the number of thetas sampled,

burn	M	N	inter	Met	α	% succ.
100	100	100	50	[30 15]	10	82.7
200	100	100	100	[5 2]	0.1	81.2
1000	200	100	50	[2 2]	1.0	78.4
1000	200	100	100	[1 1]	1.0	79.0
100	100	50	50	[1 1]	1.0	76.3

TABLE 2. Summary of Simulations

N the observed sample size, inter the number of discarded sweeps between samples, Met is the number of metropolis step corrections for the root node and for the children nodes, α is the parameter of the entropic prior and finally, % succ. is the percentage of correct classifications on 1000 random tests.

Notice that the metropolis corrections seem to help but they slow down the sampler. Notice also the drop in performance when the sample size becomes 50.

These results show the adequacy of the entropic sampler for the classification task. However, the naïve bayes DAG is not competitive with DAGs containing more realistic structure for this problem. A simulated annealing search over the space of DAGs produces structures showing over 84% success rate in the more difficult task of classification with 4 (not just 2) categories of credit card.

5. Entropic Prior for Mixtures of Gaussians

The need for flexible, informative, proper priors for mixtures has been in the statistician's wish list for a long time (e.g. see [**?**]). In this section we derive, from first principles, the entropic prior for a finite mixture of gaussians. This seems to be the first informative prior for mixtures, derivable from an objective principle. The straight forward application of (6) produces a prior that on the one hand is remarkably close to the conjugate prior that has been shown most successful in simulations, and on the other hand, departs from it in a way that has always thought to be desirable but for which there was no known way to implement.

5.1. THE MODEL

We consider a finite mixture of k univariate gaussians with vector of parameters $\theta = (\mu, \sigma, \omega)$ where $\mu \in \mathbb{R}^k$ is the vector of k means, $\sigma \in \mathbb{R}_+^k$ is the vector of k standard deviations and $\omega \in \Delta^{k-1}$ is the mixing probability vector in the $(k-1)$-dimensional simplex Δ^{k-1}. We use the standard missing data model for mixtures, i.e., we assume the data is (x, z) has joint density, for $x \in \mathbb{R}$ and $z \in \{1, 2, \ldots, k\}$ given by,

$$f(x, z|\theta) = \omega_z N(x; \mu_z, \sigma_z) \tag{72}$$

where $N(x; a, b)$ denotes the density of the normal distribution with mean a and standard deviation b. The label z is assumed to be missing from the data so that

the marginal density of x has the desired mixture form,

$$f(x|\theta) = \sum_{j=1}^{k} \omega_j N(x; \mu_j, \sigma_j) \tag{73}$$

The trick is to compute the prior on the complete (x, z) likelihood to disentangle the expression for the entropy.

5.2. ENTROPY

Let $\theta^o = (m, s, \omega^o)$ be the initial guess for θ. The Kullback number between two distributions (72) with parameters θ and θ^o is,

$$I(\theta : \theta^o) = E_\theta \left(\log \frac{\omega_z N(x; \mu_z, \sigma_z)}{\omega_z^o N(x; m_z, s_z)} \right) \tag{74}$$

Computing the expectation by first conditioning on z we obtain,

$$\begin{aligned} I(\theta : \theta^o) &= \sum_{j=1}^{k} \omega_j \left\{ I(N(\mu_j, \sigma_j^2) : N(m_j, s_j^2)) + \log \frac{\omega_j}{\omega_j^o} \right\} \\ &= \sum_{j=1}^{k} \omega_j \left\{ \log \frac{s_j}{\sigma_j} + \frac{(\mu_j - m_j)^2}{2s_j^2} + \frac{\sigma_j^2}{2s_j^2} - \frac{1}{2} + \log \frac{\omega_j}{\omega_j^o} \right\} \end{aligned} \tag{75}$$

Notice that since $\sum_{j=1}^{k} \omega_j = 1$ we can take the $1/2$ outside the sum and it will get absorbed into the proportionality constant for the entropic prior.

5.3. VOLUME ELEMENT

Using (24) we can immediately obtain from (75) the entries of the Fisher matrix. The matrix is clearly block diagonal with gaussian blocks for the (μ, σ) parameters and a multinomial block for the ω parameters. From the standard volume elements for gaussians and multinomials we can write the full volume element as,

$$g^{1/2}(\theta)\ d\theta = \frac{d\mu\ d\sigma\ d\omega}{\left(\prod_{j=1}^{k} \sigma_j^2\right) \left(\prod_{j=1}^{k} {\omega_j}^{1/2}\right)} \tag{76}$$

where we are abusing the notation a bit since $d\omega$ must be understood as $\prod_{j=1}^{k-1} d\omega_j$ so that $\omega \in \Delta^{k-1}$.

5.4. ENTROPIC PRIOR

Just multiply $e^{-\alpha I(\theta:\theta^o)}$ with (76) to get,

$$\pi(\theta|\alpha, \theta^o) \quad \propto \quad \prod_{j=1}^{k} \exp \left\{ -\alpha \omega_j \frac{(\mu_j - m_j)^2}{2s_j^2} \right\}.$$

$$\prod_{j=1}^{k} \left(\sigma_j^2\right)^{\frac{\alpha\omega_j}{2}-1} \exp\left\{-\frac{\alpha\omega_j}{2s_j^2}\ \sigma_j^2\right\} \cdot$$
$$\prod_{j=1}^{k} \left(\frac{\omega_j^o}{s_j}\right)^{\alpha\omega_j} {\omega_j}^{-\alpha\omega_j-1/2} \tag{77}$$

This is a remarkable result. Equation (77) says that conditional on ω all the components of μ and σ are independent and independent of each other. Moreover,

$$\mu_j|\omega \ \rightsquigarrow\ N\left(m_j, \frac{s_j^2}{\alpha\omega_j}\right) \tag{78}$$

$$\sigma_j^2|\omega \ \rightsquigarrow\ \text{Gamma}\left(\frac{\alpha\omega_j-1}{2}, \frac{\alpha\omega_j}{2s_j^2}\right) \tag{79}$$

where to obtain (79) we have used the change of variables $v = \sigma_j^2$ that produces the jacobian $v^{-1/2}$. The joint marginal density of ω is obtained by integrating (77) over μ and σ coordinates obtaining, up to a proportionality constant that,

$$\begin{aligned}\omega \ &\rightsquigarrow\ \prod_{j=1}^{k} \left\{\frac{s_j}{{\omega_j}^{1/2}} \cdot \frac{\Gamma((\alpha\omega_j-1)/2)}{(\alpha\omega_j/s_j^2)^{(\alpha\omega_j-1)/2}} \cdot \left(\frac{\omega_j^o}{s_j}\right)^{\alpha\omega_j} {\omega_j}^{-\alpha\omega_j-1/2}\right\} \\ &\rightsquigarrow\ \prod_{j=1}^{k} \frac{(\omega_j^o)^{\alpha\omega_j}\Gamma((\alpha\omega_j-1)/2)}{{\omega_j}^{(3\alpha\omega_j+1)/2}}\end{aligned} \tag{80}$$

5.5. POSTERIOR

Let $x^n = (x_1, \ldots, x_n)$ be the observed data and let z^n be the missing labels. As usual we shake the bayesian wand to obtain,

$$\begin{aligned}\pi(\theta, z^n|x^n, \alpha, \theta^o) \ &\propto\ f(x^n|\theta, z^n) f(z^n|\theta)\pi(\theta|\alpha, \theta^o) \\ &\propto\ \left(\prod_{i=1}^{n} \frac{1}{\sigma_{z_i}} \exp\left\{\frac{-(\mu_{z_i}-x_i)^2}{2\sigma_{z_i}^2}\right\}\right)\left(\prod_{i=1}^{n} \omega_{z_i}\right)\pi(\theta|\alpha, \theta^o)\end{aligned} \tag{81}$$

For $j = 1, \ldots, k$ define $k_j \in \{1, 2, \ldots, n\}$ by,

$$k_j = |\{i : z_i = j\}| \tag{82}$$

and replacing these counts into (81) we have,

$$\pi(\theta, z^n|x^n, \alpha, \theta^o) \propto \prod_{j=1}^{k} \left\{\frac{{\omega_j}^{k_j}}{\sigma_j^{k_j}} \exp\left\{\frac{-1}{2\sigma_j^2} \sum_{i:z_i=j} (\mu_j - m_j)^2\right\}\right\} \pi(\theta|\alpha, \theta^o) \tag{83}$$

5.6. GIBBS SAMPLER

Inference is done by sampling (θ, z^n) vectors from the posterior (83). To sample from (83) we use Gibbs sampling, i.e. we cycle over the full conditionals for each of the parameters. Let us use the notation $|\ldots$ to mean given all the other parameters and the data. Here are the distributions for each of the terms:

5.6.1. *Conditional for z^n*

When the vector of mixing probabilities ω is given the joint distribution of z^n are independent multinomials with ω as the parameter and independent of everything else. Thus, for $i = 1, 2, \ldots, n$

$$z_i|\ldots \rightsquigarrow \text{Multi}(\omega_1, \omega_2, \ldots, \omega_k) \tag{84}$$

5.6.2. *Conditional for μ*

Here again we have the classic problem of computing the posterior distribution for the mean of a gaussian given k_j independent gaussian observations when the prior is the conjugate gaussian. Looking at the first term of (77) and the right hand side of (83) we get,

$$\mu_j|\ldots \rightsquigarrow N(a_j, b_j^2) \tag{85}$$

where,

$$a_j = \frac{\dfrac{1}{\sigma_j^2}\displaystyle\sum_{i:z_i=j} x_i + \frac{\alpha\omega_j}{s_j^2}m_j}{\dfrac{k_j}{\sigma_j^2} + \dfrac{\alpha\omega_j}{s_j^2}} \tag{86}$$

and

$$\frac{1}{b_j^2} = \frac{k_j}{\sigma_j^2} + \frac{\alpha\omega_j}{s_j^2} \tag{87}$$

5.6.3. *Conditional for σ*

Collecting all the factors with σ_j from (83) and the second term from (77) we obtain,

$$\sigma_j|\ldots \rightsquigarrow (\sigma_j^2)^{\frac{1}{2}(\alpha\omega_j - k_j)-1} \exp\left\{\frac{-1}{2\sigma_j^2}\sum_{i:z_i=j}(\mu_j - m_j)^2 - \frac{\alpha\omega_j}{2s_j^2}\sigma_j^2\right\} \tag{88}$$

Now let $v = \sigma_j^2$, then

$$f_v(v) = f_{\sigma_j}(\sqrt{v})\frac{1}{2}v^{-1/2} \tag{89}$$

Using (89) with (88) we get,

$$v = \sigma_j^2|\ldots \rightsquigarrow v^{-a-1}\exp\left\{\frac{-c}{v} - bv\right\} \tag{90}$$

where,

$$a = \frac{1}{2}(k_j + 1 - \alpha\omega_j) \quad (91)$$

$$b = \frac{\alpha\omega_j}{s_j^2} \quad (92)$$

$$c = \frac{1}{2}\sum_{i:z_i=j}(\mu_j - x_i)^2 \quad (93)$$

We can obtain a useful alternative to (90) by doing $u = 1/v$ so that

$$f_u(u) = f_v(u^{-1})u^{-2}$$

and we get,

$$u = \sigma_j^{-2}|\ldots \rightsquigarrow u^{a-1}\exp\left\{\frac{-b}{u} - cu\right\} \quad (94)$$

where a, b and c are given by (91), (92), and (93) as before.

The distributions (90) and (94) are instances of the so called *Generalized Inverse Gaussian* (or GIG for short, see [**?**]) distribution. The GIG distribution was first introduced in relation to hyperbolic distributions in [**?**]. It can be shown that,

$$\int_0^\infty u^{a-1}\exp\left\{\frac{-b}{u} - cu\right\}\, du = 2\left(\frac{b}{c}\right)^{\frac{a}{2}}\text{BesselK}(a, 2\sqrt{bc}) \quad (95)$$

where the BesselK(a, x) is the modified Bessel function of the third kind. It is the solution to the differential equation,

$$x^2y'' + xy' - (x^2 + a^2)y = 0 \quad (96)$$

Thus, (90) and (94) are proper provided that $b > 0$ and $c > 0$. When either $b = 0$ or $c = 0$ (but not both) one of the two becomes a Gamma. As it is indicated in [**?**] the good news about GIGs is that they are log concave and there are universal algorithms for generating them. The problem is that the standard off the shelve algorithm for log concave densities requires the evaluation of the normalization constant, which in this case is too expensive, since it involves evaluating BesselK. The following Gamma approximation provides a solution to this problem.

5.6.4. *Gamma Approximation to GIG*

By computer algebra it is possible to find the parameters of a Gamma that best fit a given GIG. Let us use the notation, for $\alpha > 0$ and $\beta > 0$,

$$\Gamma(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\ e^{-\beta x} \text{ for } x > 0 \quad (97)$$

and let, for $a > 0, b > 0$ and $c > 0$,

$$G(x; a, b, c) = \frac{1}{Z}x^{a-1}\ \exp\left\{\frac{-b}{x} - cx\right\} \text{ for } x > 0 \quad (98)$$

where Z is the normalization constant given by the right hand side of (95). We summarize the findings in the next theorem.

Theorem 2 *The best second order* $\Gamma(x;\alpha^*,\beta^*)$ *approximation to* $G(x;a,b,c)$ *is when,*

$$\alpha^* = a\left[1+\frac{4bc}{\lambda}\right] \tag{99}$$

$$\beta^* = c\left[1+\frac{4bc}{\rho}\right] \tag{100}$$

where,

$$\lambda = a-1+E \tag{101}$$

$$\rho = (a-1)\lambda \tag{102}$$

$$E = \sqrt{(a-1)^2+4bc} \tag{103}$$

Proof Here is a summary of what was found with MAPLE. The function $G(x;a,b,c)$ has a single global maximum at

$$x^* = \frac{\lambda}{2c} \tag{104}$$

Expanding both log likelihoods in Taylor series about x^* we get,

$$\log\Gamma(x;\alpha,\beta) = A_0 + A_1(x-x^*) + A_2(x-x^*)^2 + o((x-x^*)^2) \tag{105}$$

$$\log G(x;a,b,c) = B_0 + 0\cdot(x-x^*) + B_2(x-x^*)^2 + o((x-x^*)^2) \tag{106}$$

The optimal parameters α^* and β^* are the solution to the system of equations,

$$A_1(\alpha,\beta) = 0 \tag{107}$$

$$A_2(\alpha,\beta) = B_2(a,b,c) \tag{108}$$

Q.E.D.

The Gamma approximation provided by theorem 2 fits the bulk of the GIG very well but the tails of the GIG are always heavier. A few metropolis iterations starting from the gamma approximation should be used to correct for the light tails.

5.6.5. *Conditional for* ω
Collecting all the factors with ω_j from (83) and all the terms from (77) we obtain,

$$\omega|\ldots \rightsquigarrow \prod_{j=1}^{k} {\omega_j}^{\alpha_j-1}\ e^{-\beta_j\omega_j} \tag{109}$$

where,

$$\alpha_j = k_j - \alpha\omega_j + 1/2 \approx k_j + 1/2 \tag{110}$$

and,

$$\beta_j = \frac{\alpha}{2}\left[\left(\frac{\mu_j-m_j}{s_j}\right)^2 + \left(\frac{\sigma_j}{s_j}\right)^2 - \log\left(\frac{\sigma_j}{s_j}\right)^2 + 2\log\frac{1}{\omega_j^o}\right] \tag{111}$$

Notice that $\beta_j > 0$ and we can use Lemma 1 again to find good starting approximations to be corrected with a small number of metropolis iterations.

6. Conclusions and Future Work

We have provided explicit formulas for adding objective prior information in two general classes of hypothesis spaces: Discrete probabilistic networks and mixtures of gaussians models. Many highly successful models are special cases of BBNs. A partial list lifted from [**?**] include, linkage analysis in genetics, Hidden Markov Models for speech recognition, Kalman filtering for tracking missiles, and density estimation for data compression and coding with turbocodes. It is only natural to expect improvements in the performance of these methods if there is available cogent prior information that has not been used. This is specially true in high dimensional parametric models.

I am currently investigating alternative/complementary methods to MCMC for performing approximate inference with entropic piors. These include, the variational bayes approach (see [**?**]), and the Expectation Propagation (EP) method of Minka (see [**?**]).

7. Acknowledgments

This paper was conceived during the summer of 2000 while I was visiting the data analysis group at the *Center for Interdisciplinary Plasma Science* (CIPS) [**?**]. I would like to thank Volker Dose, Rainner Fischer, Roland Preuss, Udo von Toussaint and Silvio Gori for many stimulating conversations.